Capstone Project 2: Project Proposal

**What is the problem you want to solve?**

 Diabetes is a medical condition that affects approximately 1 in 10 patients in the United States. This patient population is one with an increased risk for hospitalization and readmission. I propose to use predictive modeling from data science to help identify patients with a risk for hospital readmission.

**Who is your client and why do they care about this problem? In other words, what will your client do or decide based on your analysis that they wouldn’t have done otherwise?**

The Centers for Medicare & Medicaid Services, CMS which is part of the Department of Health and Human Services (HHS) has created many programs to improve the quality of care of patient as the healthcare system moves toward value-based care. Hospital Readmission Reduction Program ([HRRP](https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/Value-Based-Programs/HRRP/Hospital-Readmission-Reduction-Program.html)), which is one of them , reduces reimbursement to hospitals with above average readmissions. For those hospitals which are currently penalized under this program, one solution is to create interventions to provide additional assistance to patients with increased risk of readmission.

**What data are you using? How will you acquire the data?**

Data is available in UCI machine learning repository ([link](https://archive.ics.uci.edu/ml/datasets/diabetes+130-us+hospitals+for+years+1999-2008)). The data consists of over 100000 hospital admissions from patients with diabetes from 130 US hospitals between 1998 and 008.

**Briefly outline how you’ll solve this problem. Your approach may change later, but this is a good first step to get you thinking about a method and solution.**

I plan to build a model predicting readmission in Python using

data exploration ,feature engineering ,building training/validation/test samples

model selection and finally model evaluation.

The dependent variable is clearly given and so a supervised learning algorithm would be the best option. I plan to use

* Logistic regression
* K-nearest neighbors
* Stochastic gradient descent
* Naive Bayes
* Decision tree
* Random forest
* Gradient boosting classifier

Though the dependent variable is present, I would also like to see if clustering can be used to try an implementation of unsupervised learning to see if there are clear patterns of clusters evolving and plan to used the K Means to derive the clusters

**What are your deliverables? Typically, this includes code, a paper, or a slide deck.**

The deliverables for this project would be a complete presentation of the project with all the documentations, visualizations and python code on notebook. The final submission will include the introduction, the steps and methods involved in this project from the beginning to the end of the project.